Nucleic Acids Research, 2006, Vol. 34, Database issue ***D273–D276***
doi:10.1093/nar/gkj018

DMAPS: a database of multiple alignments for protein structures

Chittibabu Guda*, Lipika R. Pal and Ilya N. Shindyalov[1]

Gen*NY*sis Center for Excellence in Cancer Genomics and Department of Epidemiology and Biostatistics, University at Albany, State University of New York, One Discovery drive, Rensselaer, NY 12144-3456,USA and [1]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA

Received August 15, 2005; Accepted September 17, 2005

ABSTRACT

The database of multiple alignments for protein structures (DMAPS) provides instant access to pre-computed multiple structure alignments for all protein structure families in the Protein Data Bank (PDB). Protein structure families have been obtained from four distinct classification methods including SCOP, CATH, ENZYME and CE, and multiple structure alignments have been built for all families containing at least three members, using CE-MC software. Currently, multiple structure alignments are available for 3050 SCOP-, 3087 CATH-, 664 ENZYME- and 1707 CE-based families. A web-based query system has been developed to retrieve multiple alignments for these families using the PDB chain ID of any member of a family. Multiple alignments can be viewed or downloaded in six different formats, including JOY/html, TEXT, FASTA, PDB (superimposed coordinates), JOY/postscript and JOY/rtf. DMAPS is accessible online at http://bioinformatics.albany.edu/~dmaps.

INTRODUCTION

The number of protein structures deposited into the Protein Data Bank (PDB) has more than doubled in the past 5 years (http://www.rcsb.org/pdb). This trend is partly due to the advances in high-throughput crystallography techniques (1) as well as the Protein Structure Initiative (http://www.nigms.nih.gov/psi) funded by the NIGMS (2). To classify and analyze such rapidly growing number of protein structures, several structure family databases exist (3–8) where each database uses a different set of criteria for classification. For example, SCOP (Structural Classification of Proteins) database (3) classifies proteins of known structure based on their evolutionary and structural relationship, whereas CATH (4) uses a hierarchical order at four major levels, i.e. Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). Structure-based alignment of the members of these families is the key for identifying the structurally conserved domains and patterns. Protein structure comparison servers, such as FSSP (Families of Structurally Similar Proteins) (5) and CE (Combinatorial Extension) (6) provide alignments between structural neighbors in the PDB based on exhaustive all-to-all pairwise comparisons. Other resources such as HOMSTRAD (Homologous Structural Alignment Database) provide multiple structure alignments for a selected set of families that do not include all PDB chains (8). Nevertheless, there are no web resources that offer multiple structure alignments for all PDB chains classified by different criteria.

We developed the CE-MC algorithm for the alignment of multiple protein structures based on Monte Carlo (MC) optimization (9,10), CE-MC web server for computing alignments online and CE-MC standalone software for local installation (11). Computing multiple structure alignments for larger families is very time-consuming; hence, here we present DMAPS, a database of multiple alignments for protein structures that provides instant access to pre-computed multiple structure alignments for all possible protein chains in the PDB as classified by SCOP, CATH, CE and ENZYME databases.

DATABASE DEVELOPMENT

CE-MC algorithm

CE-MC program works based on two independent algorithms, i.e. CE and MC optimization. CE algorithm (6) generates all-against-all pairwise alignments for all chains in a given family and the chain with the highest average *Z*-score (over the alignments to the other family members) is selected as the ''master'' chain, while the remaining chains in the family are designated as ''slaves''. As a zero-approximation, MC algorithm assembles the seed alignment in a ''master-slave''

*To whom correspondence should be addressed. Tel: +1 518 591 7155; Fax: +1 518 591 7201; Email: cguda@albany.edu

fashion and performs the optimization process as described in Guda *et al*. (9). In a nutshell, global optimization of multiple alignments is accomplished by random and iterative exploration of the search space with occasional excursions into the non-optimal territory, until the optimization converges.

Datasets for protein structure families in PDB

Protein structure families have been obtained from four different sources which use non-overlapping criteria for classification. These include SCOP (12), CATH (13), ENZYME (14) and CE classifications (15). Although CATH and SCOP use evolutionary and topological features for classification, ENZYME classification is based on the functional similarities (with identical EC number) and CE classification is purely based on the substructure similarities using C-α coordinate distances. SCOP database (release 1.67) includes 24 037 PDB entries corresponding to 50 285 individual chains, excluding non-protein entries and theoretical models. Each entry in the SCOP database is classified in the hierarchical order of Class → Fold → Superfamily → Family → Protein domain → Species → PDB entry. DMAPS database contains multiple structure alignments for all PDB entries at each protein domain level. CATH database (release 2.6.0) considers NMR structures and those crystal structures with resolution better than 3.0 Å, excluding all non-protein and ''C-α only'' structures. CATH uses a hierarchical classification of protein domain structures as Class (C), Architecture (A), Topology (T), Homologous Superfamily (H) and Sequence families (S) with sequence identity >35%. DMAPS provides alignments at the S-level for 22 478 PDB entries corresponding to 45 414 individual chains. ENZYME database (release 37) classifies enzymes with known structures in the PDB based on their enzyme classification (EC) numbers reflecting pure functional similarities. This release contains 15 106 PDB entries that include only NMR and crystal structures excluding theoretical models. The last dataset of PDB families was obtained from CE classification based on the substructure homologies (15). In this case, PDB chains were aligned all-against-all in a pairwise fashion and neighbors with *Z*-score >4.0 and RMSD (root mean squared deviation) <3.0 Å have been assembled into clusters of common substructures. In all the datasets above, if the size of a family exceeds 50, we selected chains only from unique PDB entries to reduce redundancy. After the first filtering, if the number still exceeds 50, only the first 50 members were used for computing multiple structure alignments. However, the filtered chains have been indexed to their original family number to facilitate retrieval of multiple alignments for corresponding family.

Building multiple alignments

Multiple structure alignments have been generated for all structure families described above, using our CE-MC standalone software (11) (available for download from the CE-MC web server at http://bioinformatics.albany.edu/~cemc). CE-MC source code has been compiled and run on a Dell four-node cluster with dual Xeon 64-bit, 3.2 GHz CPUs per node, running on Fedora Core 3 Linux OS. Automated multiple alignments have been generated for all structural families containing at least three member chains. CE-MC program automatically filters out those protein chains with average *Z*-score < 4.0 to include only structurally homologous proteins in the multiple alignments. CE-MC output comes in text format by default and several intermediate scripts have been developed to reformat the alignment in other user-friendly formats including FASTA and JOY-based formats. JOY program (16) uses 3D coordinates from PDB files to calculate secondary structural and local environmental features and displays the same in a color-coded format in the sequence alignment.

DATABASE ACCESS

Currently, multiple structure alignments are available for 3050 SCOP-, 3087 CATH-, 664 ENZYME- and 1707 CE-based families. Under each classification, PDB chains belonging to one or more families are indexed to enable easy retrieval. In cases where a PDB chain belongs to multiple families, multiple alignments for all families can be retrieved. Queries may be made by entering any PDB chain ID such as 1CDK:A for multiple chain entries or 1AK1:_ for single chain entries. A CGI script finds corresponding family or families for any given PDB chain ID and displays hyperlinks to view or download the results. Hyperlinks to the source pages of family classifications and PDB files are also displayed in the results page (Figure 1). Multiple alignments can be downloaded in six different formats, i.e. JOY-html, Text, FASTA, PDB, JOY-postscript and JOY-rich text. The file in PDB format contains superimposed coordinates for multiple alignments that could be directly imported into 3D visualization programs such as PyMOL (http://pymol.sourceforge.net).

DISCUSSION

In the DMAPS database, we generated multiple alignments for proteins belonging to protein domain level in SCOP and sequence family level in CATH. Typically, these classification levels consist of protein domains that share very similar sequences and structures and hence the alignments generated in these two cases are usually unambiguous. However, alignments built based on the ENZYME classification have functional significance whereas those from CE classification are based on the substructure similarities. DMAPS database has been developed to provide quick access to multiple structure alignments for all protein families in the PDB. Building structure alignments using the CE-MC algorithm is time-consuming as the computational time grows quadratic to the number of chains in the family. Hence, alignments have been pre-computed, indexed and stored in flat files for ready access from the web interface. As shown in Table 1, automated alignments have been successfully generated for ~80% of the CATH, SCOP and ENZYME families and 100% of the CE families containing at least three members. For some families, we could not compute the alignments and similarly for some, JOY formatting could not be carried out owing to unidentified bugs in the program code. We are in the process of fixing these bugs and in the near future, we hope to update the DMAPS database with multiple alignments for all structural families in PDB containing at least three members.

Current structure databases, such as HOMSTRAD (8), DALI (17), PALI (18) and PASS2 (19), also provide structural

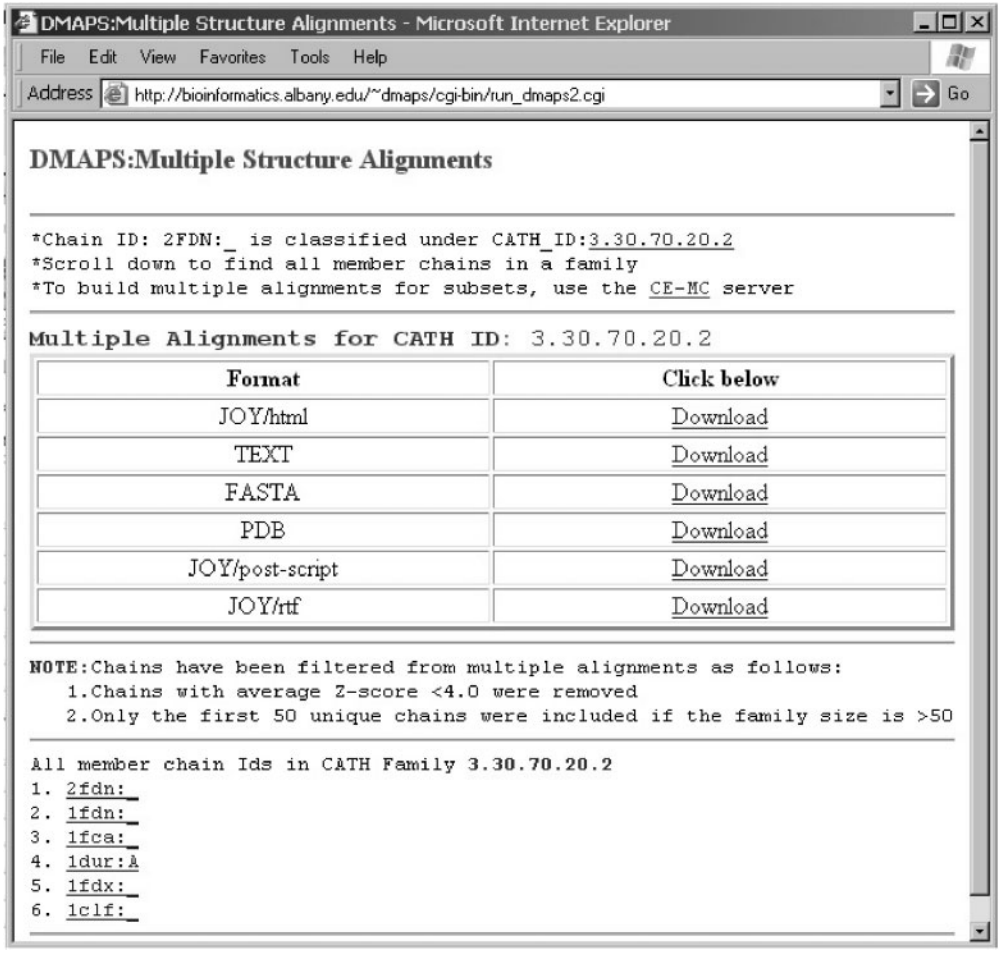

Figure 1. A screen shot of the DMAPS query results.

Table 1. Multiple structure alignments for PDB chains using four classification methods

Classification methods	Number of families	Families with three or more members	Multiple alignments built
SCOP	6475	3788	3050
CATH	6003	3842	3087
ENZYME	1039	843	664
CE	4207	1708	1707

alignments for PDB entries, but several limitations exist in their usability compared with the DMAPS database. For instance, HOMSTRAD provides multiple alignments only to representative members of a family while PASS2 covers only 628 multimember families. However, PALI builds multiple alignments progressively from pairwise alignments whereas DALI offers only pairwise alignments based on all-against-all comparison of 3D coordinates in the PDB. The unique features of DMAPS database are as follows: it offers instant access to multiple structure alignments for thousands of protein families classified by several functional and structural criteria, and the alignments are available in a variety of formats for further use. The DMAPS database will be useful for an array of research projects in structural bioinformatics, such as building structure-based family profiles and patterns, improving the quality of sequence alignments and finding remote homologues.

ACKNOWLEDGEMENTS

The authors are thankful to Prof. Philip Bourne and Dr Eric Scheeff for their role in the development of CE-MC program and Dr Kenji Mizuguchi for providing JOY binaries. This work has been supported by the start-up funds to C.G. from the State University of Albany at New York (SUNY). Funding to pay the Open Access publication charges for this article was provided by the SUNY start-up funds.

Conflict of interest statement. None declared.

REFERENCES

1. Lesley,S.A., Kuhn,P., Godzik,A., Deacon,A.M., Mathews,I., Kreusch,A., Spraggon,G., Klock,H.E., McMullan,D., Shin,T. *et al*. (2002) Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
2. Stevens,R.C., Yokoyama,S. and Wilson,I.A. (2001) Global efforts in structural genomics. *Science*, **294**, 89–92.
3. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements

integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.
4. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al*. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*., **33**, D247–D251.
5. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.
6. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
7. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol*., **6**, 377–385.
8. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci*., **7**, 2469–2471.
9. Guda,C., Scheeff,E.D., Bourne,P.E. and Shindyalov,I.N. (2001) A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization. *Pac. Symp. Biocomput*., 275–286.
10. Guda,C., Scheeff,E.D., Bourne,P.E. and Shindyalov,I.N. (2002) Comparative analysis of protein structure: new concepts and approaches for multiple structure alignment. In Tsigelny,I.F. (ed.), *Protein Structure Prediction: Bioinformtic Approach*. Int'ernational University Line, La Jolla, CA, pp. 451–459.
11. Guda,C., Lu,S., Sheeff,E.D., Bourne,P.E. and Shindyalov,I.N. (2004) CE-MC: a multiple protein structure alignment server. *Nucleic Acids Res*., **32**, W100–W103.
12. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
13. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
14. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
15. Shindyalov,I.N. and Bourne,P.E. (2000) An alternative view of protein fold space. *Proteins*, **38**, 247–260.
16. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998) JOY: protein sequence–structure representation and analysis. *Bioinformatics*, **14**, 617–623.
17. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res*., **26**, 316–319.
18. Gowri,V.S., Pandit,S.B., Karthik,P.S., Srinivasan,N. and Balaji,S. (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res*., **31**, 486–488.
19. Bhaduri,A., Pugalenthi,G. and Sowdhamini,R. (2004) PASS2: an automated database of protein alignments organised as structural families. *BMC Bioinformatics*, **5**, 35.